

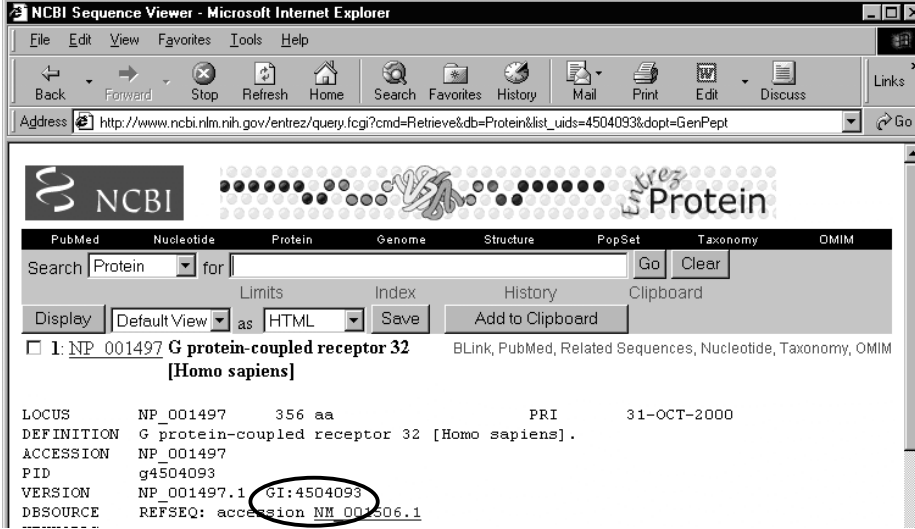
Phylogenomic analysis

Kimmen Sjolander
Department of Bioengineering
University of California, Berkeley
kimmen@uclink.berkeley.edu
URL: phylogenomics.berkeley.edu
URL: alumni.cse.ucsc.edu/~kimmen/

Homolog detection is just the first step

The phylogenetic context is critical

Example 1: Orphan GPCR classification



NCBI Sequence Viewer - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&list_uids=4504093&dopt=GenPept

NCBI Protein

Search Protein for [] Go Clear

Display Default View as HTML Save Add to Clipboard

☐ 1: NP_001497 G protein-coupled receptor 32 [Homo sapiens]

LOCUS NP_001497 356 aa PRI 31-OCT-2000

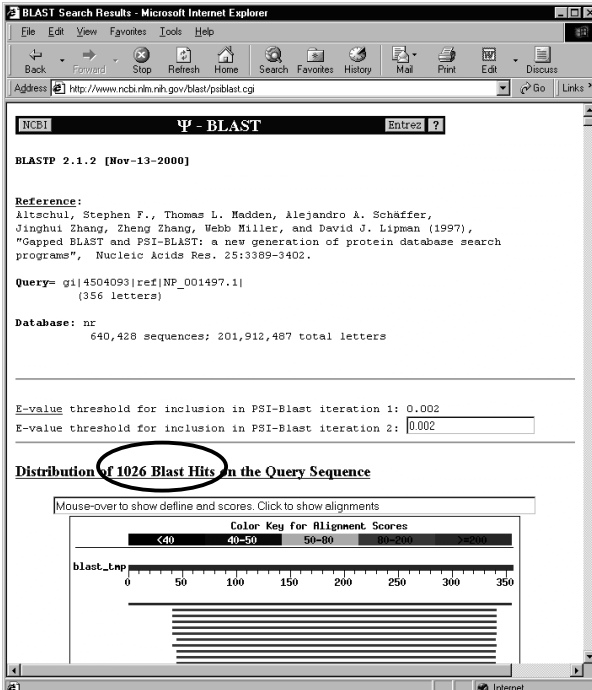
DEFINITION G protein-coupled receptor 32 [Homo sapiens].

ACCESSION NP_001497

PID g4504093

VERSION NP_001497.1 GI:4504093

DBSOURCE REFSEQ: accession NM_003506.1



BLAST Search Results - Microsoft Internet Explorer

Address: <http://www.ncbi.nlm.nih.gov/blast/psblast.cgi>

NCBI Ψ-BLAST

BLASTP 2.1.2 [Nov-13-2000]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= g14504093|ref|NP_001497.1
(356 letters)

Database: nr
640,428 sequences; 201,912,487 total letters

E-value threshold for inclusion in PSI-Blast iteration 1: 0.002
E-value threshold for inclusion in PSI-Blast iteration 2: 0.002

Distribution of 1026 Blast Hits on the Query Sequence

Mouse-over to show define and scores. Click to show alignments

Color Key for Alignment Scores

Score Range	Color
<40	Black
40-50	Dark Grey
50-60	Medium Grey
60-70	Light Grey
70-80	White

blast_top

0 50 100 150 200 250 300 350

Step 1:
Run
BLAST

BLAST Search Results - Microsoft Internet Explorer

Address: <http://www.ncbi.nlm.nih.gov/blast/pblast.cgi>

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:

	Score	E Value
ref NP_001497.1 G protein-coupled receptor 32 [Homo sapiens]...	574	0.0
sp P79190 FML1_MACMU FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi 1...	199	6e-50
ref NP_001453.1 formyl peptide receptor-like 1; lipoxin A4 r...	198	9e-50
gb AAAS8481.1 (M76672) FMLP-related receptor II [Homo sapiens]	198	1e-49
sp P79177 FML1_GORGO FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi 1...	196	4e-49
sp P79242 FML1_PANTR FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi 1...	195	7e-49
sp P79243 FML2_PANTR N-FORMYL PEPTIDE RECEPTOR-LIKE 2 RECEPTO...	195	7e-49
sp P79236 FML1_PONPY FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi 1...	194	1e-48
ref XP_009373.1 formyl peptide receptor-like 2 [Homo sapiens]...	191	1e-47
sp P79237 FML2_PONPY N-FORMYL PEPTIDE RECEPTOR-LIKE 2 RECEPTO...	191	1e-47
ref NP_020201.2 formyl peptide receptor-like 2 [Homo sapiens]...	190	3e-47
ref NP_032066.1 formyl peptide receptor, related sequence 3 ...	187	1e-46
sp P79178 FML2_GORGO FMLP-RELATED RECEPTOR II (FMLP-R-II) >gi...	187	2e-46
sp P79191 FML2_MACMU N-FORMYL PEPTIDE RECEPTOR-LIKE 2 RECEPTO...	181	1e-44
ref NP_032065.1 formyl peptide receptor, related sequence 2 ...	181	2e-44
ref XP_009375.1 formyl peptide receptor 1 [Homo sapiens]	178	1e-43
sp P79189 FMLR_MACMU FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...	176	3e-43
pir I42002 N-formyl peptide receptor - human >gi 182740 gb A...	176	4e-43
sp P21462 FMLR_HUMAN FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...	176	6e-43
ref NP_020200.1 formyl peptide receptor 1 [Homo sapiens] >gi...	175	6e-43
ref NP_032067.1 formyl peptide receptor, related sequence 4 ...	175	7e-43
sp P79241 FMLR_PANTR FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...	173	3e-42
ref NP_038549.1 formyl peptide receptor 1 [Mus musculus] >gi...	173	3e-42
sp Q05394 FMLR_RABIT FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...	173	3e-42
sp P79235 FMLR_PONPY FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...	173	3e-42
gb AA36362.1 (M37128) N-formylpeptide receptor FMLP-R98 [Ho...	173	4e-42
ref NP_032068.1 formyl peptide receptor-like 1 [Mus musculus]...	172	5e-42
sp P79176 FMLR_GORGO FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...	171	1e-41

Is the query an FMLP receptor?

BLAST Search Results - Microsoft Internet Explorer

Address: <http://www.ncbi.nlm.nih.gov/blast/pblast.cgi>

ref NP_032064.1 formyl peptide receptor, related sequence 1 ...	170	3e-41
ref NP_004063.1 chemokine-like receptor 1 [Homo sapiens] >gi...	157	2e-37
sp Q99788 CML1_HUMAN CHEMOKINE RECEPTOR-LIKE 1 (G-PROTEIN COU...	157	2e-37
sp P97468 CML1_MOUSE CHEMOKINE RECEPTOR-LIKE 1 (G-PROTEIN COU...	156	5e-37
ref NP_071554.1 G-protein coupled chemoattractant-like recep...	144	2e-33
sp Q97664 GPR1_MACMU PROBABLE G PROTEIN-COUPLED RECEPTOR GPR1...	131	1e-29
ref XP_006864.1 chemokine-like receptor 1 [Homo sapiens]	125	1e-27
sp Q88416 GPRX_MOUSE PROBABLE G PROTEIN-COUPLED RECEPTOR GPR3...	123	4e-27
ref XP_002667.1 TR00065073_p [Homo sapiens]	121	2e-26
sp Q97093.1 g-protein coupled receptor 1 [Rattus sp.] >gi...	119	8e-26
ref NP_005270.1 G protein-coupled receptor 1 [Homo sapiens] ...	118	1e-25
sp P79175 C5AR_GORGO C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...	115	8e-25
sp P97520 C5AR_RAT C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (C5...	115	8e-25
gb AAD21055.1 (AF118265) orphan G protein-coupled receptor G...	115	9e-25
ref I1705295A anaphylatoxin C5a chemotactic receptor [Homo sa...	114	1e-24
ref NP_001727.1 complement component 5 receptor 1 (C5a ligan...	114	1e-24
gb AAC36503.1 (U86378) anaphylatoxin C3a receptor [Cavia por...	114	2e-24
sp Q88680 C3AR_CAVPO C3A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...	113	3e-24
sp P79240 C5AR_PANTR C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...	112	5e-24
ref NP_004769.1 G protein-coupled receptor 44; chemoattracta...	112	6e-24
sp P79234 C5AR_PONPY C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...	112	8e-24
ref XP_006046.1 chemokine (C-C motif) receptor 4 [Homo sapiens]	112	8e-24
gb AAF13030.1 AF068680_1 (AF068680) anaphylatoxin C5a recepto...	111	2e-23
sp P79189 C5AR_MACMU C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...	110	3e-23
ref NP_033242.1 somatostatin receptor 1 [Mus musculus] >gi 4...	109	4e-23
ref NP_036851.1 somatostatin receptor subtype 1 [Rattus norveg...	109	5e-23
gb AAG60607.1 (AF285173) opioid receptor-like protein ZFOR3 ...	109	5e-23
gb AAC50657.1 (U62027) anaphylatoxin C3a receptor [Homo sapi...	109	5e-23
sp Q16581 C3AR_HUMAN C3A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...	109	8e-23
ref NP_004045.1 complement component 3a receptor 1; compleme...	109	8e-23
ref NP_034092.1 chemoattractant receptor-homologous molecule...	107	2e-22
gb JBA04109.1 (D16829) kappa opioid receptor [Rattus norveg...	107	2e-22
sp Q55197 C3AR_RAT C3A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (C3...	107	3e-22

Or a chemokine receptor?

Or a C5A Anaphylatoxin chemotactic receptor?

Or...?

Or a novel type of receptor?

SF7-8923873	G protein-coupled receptor CSL2 [Homo sapiens]
SF0-6831553	PROBABLE G PROTEIN-COUPLED RECEPTOR GPR33
SF1-4504093	PROBABLE G PROTEIN-COUPLED RECEPTOR GPR32
SF2-9834697	ORF FPV010 Serpin gene family protein [Fowlpox virus]
SF3-9834691	ORF FPV010 Serpin gene family protein [Fowlpox virus]
SF33-PROBABLE G PROTEIN-COUPLED RECEPTOR CRTH2	ornithokinin receptor [Gallus gallus]
SF18-2660531	chemokine (C-C) receptor 9 [Mus musculus]
SF20-11024708	CXC chemokine receptor-1 [Cyprinus carpio]
SF37-PROBABLE G PROTEIN-COUPLED RECEPTOR GPR1	CXC chemokine receptor-2 [Cyprinus carpio]
SF5-3298340	
SF9-3298358	
SF35-CHEMOKINE RECEPTOR-LIKE 1 (G-PROTEIN COUPLED REC	
SF23-TYPE-2 ANGIOTENSIN II RECEPTOR (AT2)	chemokine receptor [Oncorhynchus mykiss]
SF11-2655885	
SF31-POSSIBLE GUSTATORY RECEPTOR TYPE B (PPR1 PRV	
SF29-C-X-C CHEMOKINE RECEPTOR TYPE 3 (CXC-R3) (CXC	
SF45-C3A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR	
SF46-C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR	
SF48-FMEL-LEU-PHE RECEPTOR (FMILP RECEPTOR)	
SF12-4895345	PROBABLE G PROTEIN-COUPLED RECEPTOR GPR8
SF16-4895343	PROBABLE G PROTEIN-COUPLED RECEPTOR GPR7
SF10-3122892	SOMATOSTATIN-LIKE RECEPTOR F_48010.1
SF6-9794865	somatostatin receptor type two [Carassius auratus]
SF51-SOMATOSTATIN RECEPTOR TYPE 4 (SS4R)	
SF25-SOMATOSTATIN RECEPTOR TYPE 2 (SS2R) (S	
SF21-SOMATOSTATIN RECEPTOR TYPE 3 (SS3R) (S	
SF32-SOMATOSTATIN RECEPTOR TYPE 5 (SS5R)	
SF28-NOCICEPTIN RECEPTOR (ORPHANIN FG RECEPTOR	
SF22-KAPPA-TYPE OPIOID RECEPTOR (KOR-1) (MSL-1)	
SF53-DELTA-TYPE OPIOID RECEPTOR (DOR-1) (OPIOID	
SF52-MU-TYPE OPIOID RECEPTOR (MOR-1)	
SF49-C-X-C CHEMOKINE RECEPTOR TYPE 4 (CXC-R4) (C	Mesenchyme-associated serpentine receptor [Xenopus l
SF15-1729805	angiotensin receptor [Anguilla anguilla]
SF24-PROBABLE G PROTEIN-COUPLED RECEPTOR AP	
SF4-4595688	putative chemokine receptor [Gallus gallus]
SF44-TYPE-1-LIKE ANGIOTENSIN II RECEPTOR 2 (AT1)	
SF17-3243095	
SF34-C-X-C CHEMOKINE RECEPTOR TYPE 5 (CXC-R5)	
SF40-G PROTEIN-COUPLED RECEPTOR BONZO	
SF37-C-X-C CHEMOKINE RECEPTOR TYPE 7 (CXC-R7)	

Example 2: Errors in database annotations

NCBI Sequence Viewer - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&list_uids=3941547&dopt=GenPept

NCBI Protein

Search [Protein] for [] Go Clear

Display [Default View] as [HTML] Save Add to Clipboard

1: **AAC82381** putative odorant receptor LOR3 [Lampetra fluviatilis] BLink, Related Sequences, Nucleotide, Taxonomy

LOCUS AAC82381 352 aa VRT 02-DEC-1998

DEFINITION putative odorant receptor LOR3 [Lampetra fluviatilis].

ACCESSION AAC82381

PID g3941547

VERSION AAC82381.1 GI:3941547

DBSOURCE locus AF069546 accession AF069546.1

KEYWORDS .

SOURCE European river lamprey.

ORGANISM Lampetra fluviatilis

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Hyperoartia; Petromyzontiformes; Petromyzontidae; Lampetra.

REFERENCE 1 (residues 1 to 352)

AUTHORS Berghard, A. and Dryer, L.

TITLE A novel family of ancient vertebrate odorant receptors

JOURNAL J. Neurobiol. (1998) In press

The top matching BLAST hits are also putative odorant receptors

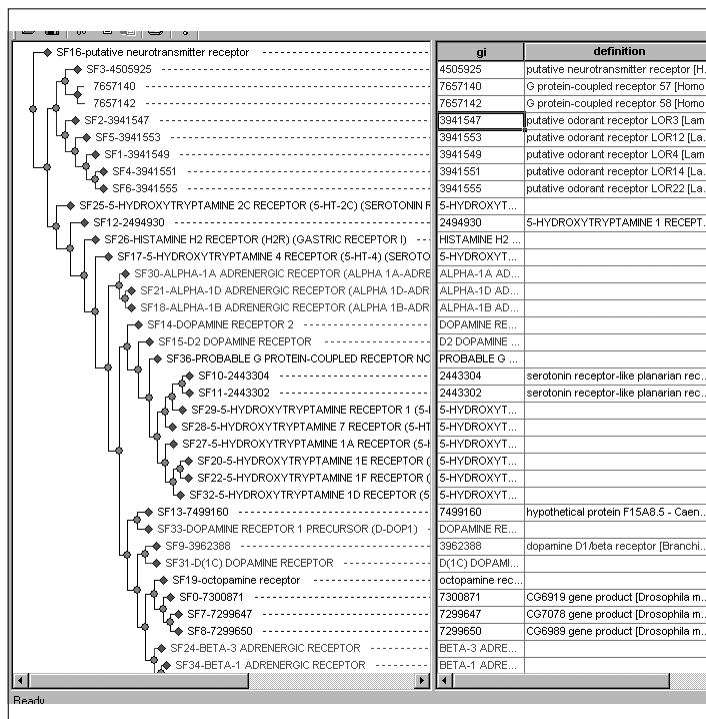
Results for RID 984268375-29090-30487 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Dis

Address http://www.ncbi.nlm.nih.gov/blast/blast.cgi

Sequences producing significant alignments:				Score	E
				(bits)	Val
gi 3941547 gb AA82381.1	(AF069546)	putative odorant recep...	633	0.0	
gi 3941549 gb AA82382.1	(AF069547)	putative odorant recep...	232	6e-60	
gi 3941553 gb AA82384.1	(AF069549)	putative odorant recep...	194	1e-48	
gi 3941551 gb AA82383.1	(AF069548)	putative odorant recep...	185	9e-46	
gi 4505925 ref NP_003958.1		putative neurotransmitter recep...	172	8e-42	
gi 7657142 ref NP_055441.1		G protein-coupled receptor 58 [...	167	2e-40	
gi 7657140 ref NP_055442.1		G protein-coupled receptor 57 [...	166	3e-40	
gi 3646424 emb CAA09599.1	(AJ011370)	serotonin 4 receptor ...	160	2e-38	
gi 3326989 emb CAA73108.1	(Y12506)	5-HT4 receptor [Homo sa...	156	3e-37	
gi 12274906 emb CAC22251.1	(AJ278982)	5-hydroxytryptamine4...	156	3e-37	



Phylogenetic analysis suggests it's more likely a Biogenic Amine GPCR

The phylogenetic context is critical

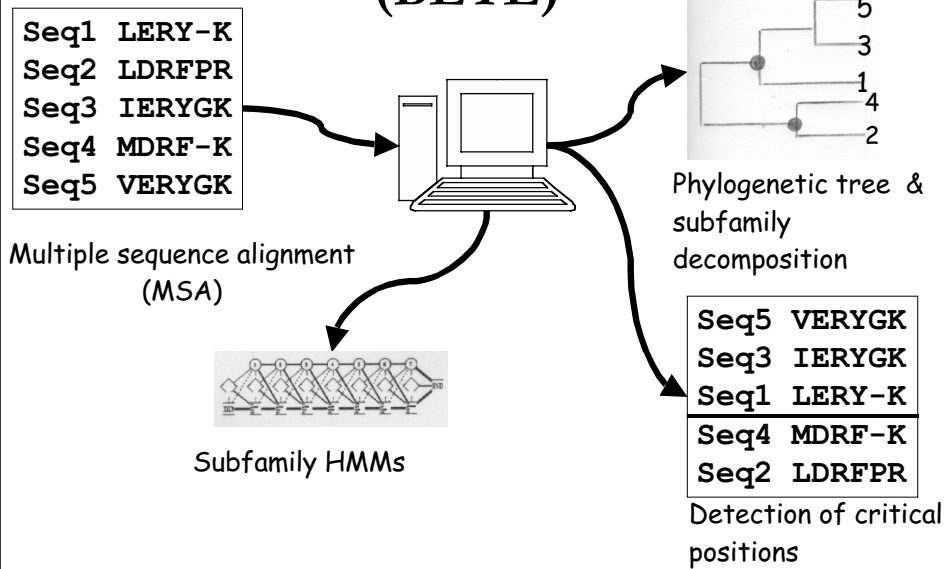
- For correctness of functional classification
- For identification of critical positions in molecules
- For structure prediction accuracy
- For detection of errors in database annotations

Process

- Gather homologs
- Generate a multiple sequence alignment
- Estimate a phylogenetic tree
- Find subfamilies
- Predict function, structure
 - Identify homologous PDB structures
 - Identify domain structure
 - Predict critical positions

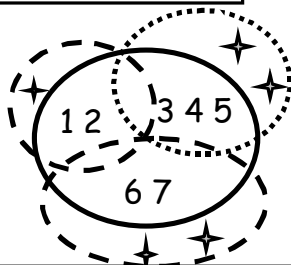
Our toolkit

Bayesian Evolutionary Tree Estimation (BETE)



How to build Subfamily HMMs (SHMMs)

1	D	S	L	F	M	K	I
2	D	S	I	F	M	K	V
3	D	T	I	W	M	K	M
4	D	T	I	W	M	K	L
5	D	T	V	W	M	K	F
6	D	T	F	R	K	K	I
7	D	T	F	R	K	K	V



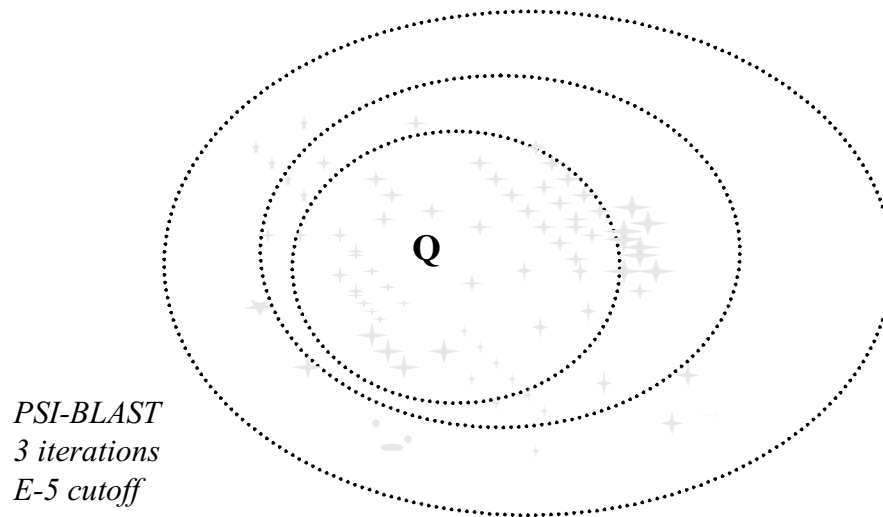
Share statistics between subfamilies where there is evidence of a common distribution.
Keep statistics separate at positions where there is evidence of ***divergent*** structure.

Improved specificity, sensitivity, alignment accuracy

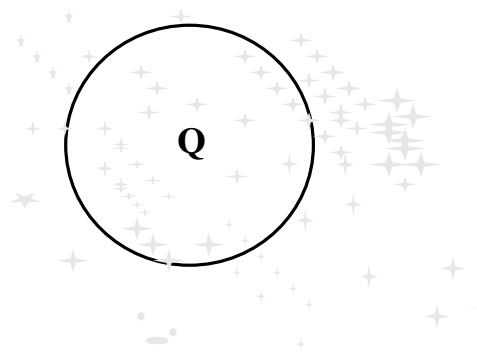
FlowerPower

Iterative clustering and alignment tool

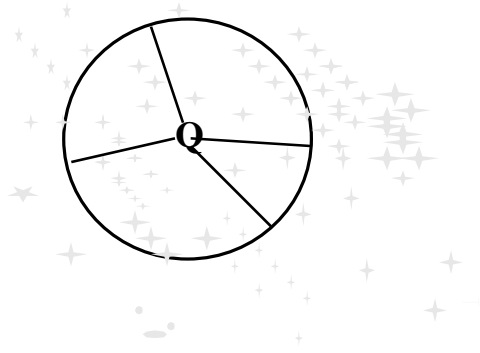
Step 1: Identify putative homologs to query sequence (Q)



Step 2: Select initial training set

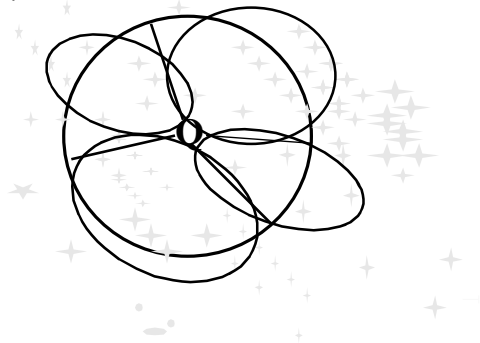


Step 3: Align initial set, identify subfamilies, and build subfamily HMMs.

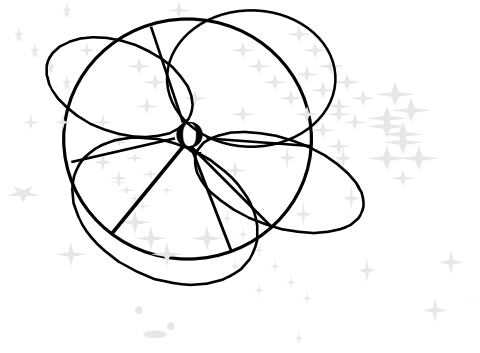


Step 4: Identify and align new homologs.

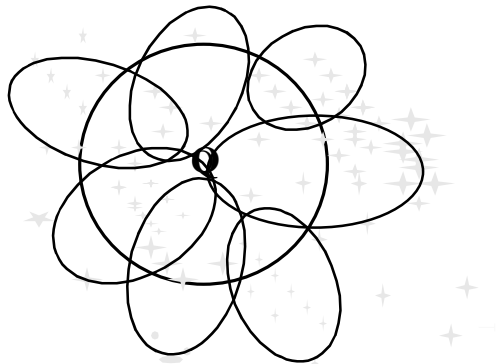
1. Search with subfamily and general HMMs.
2. Accept hits above threshold.
3. Align accepted hits to closest HMM.



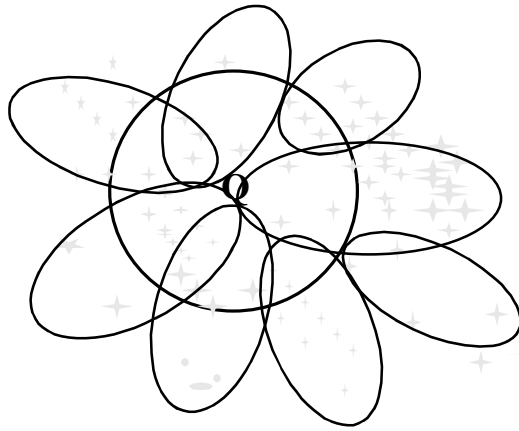
Step 5: Run BETE to identify subfamilies, and build new subfamily HMMs.



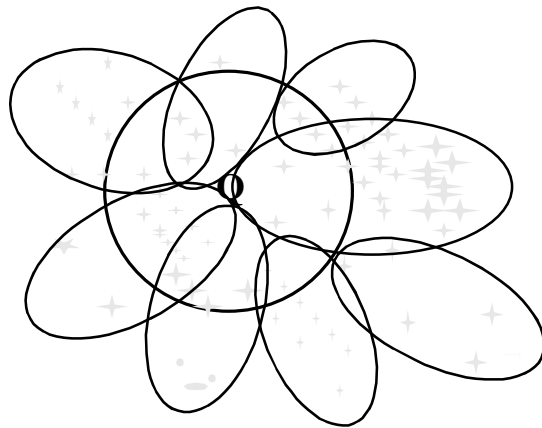
Step 6: Iterate



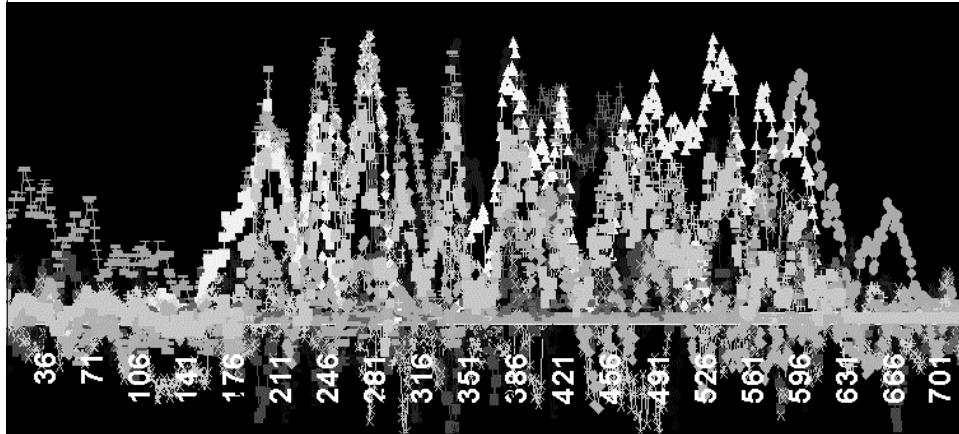
until ...



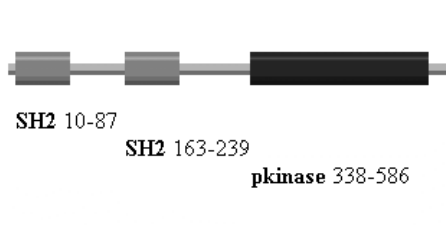
convergence.



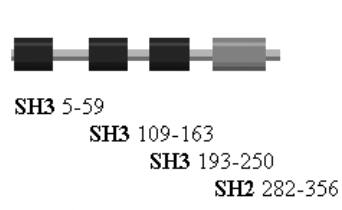
Domain Structure Analysis



Common domains are found in proteins of diverse function and structure

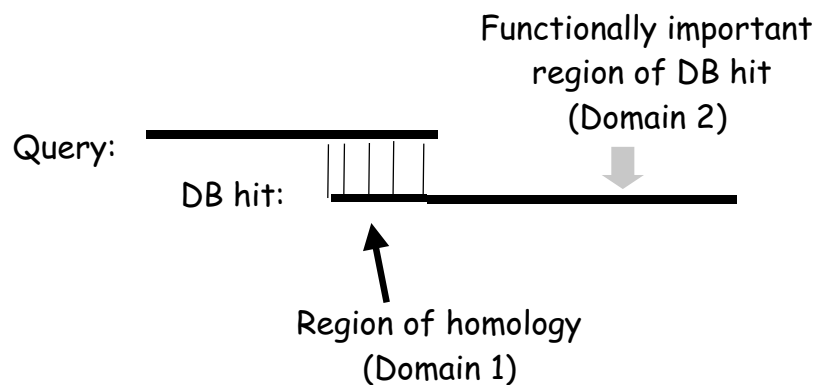


ZA70_human
(tyrosine-protein kinase)

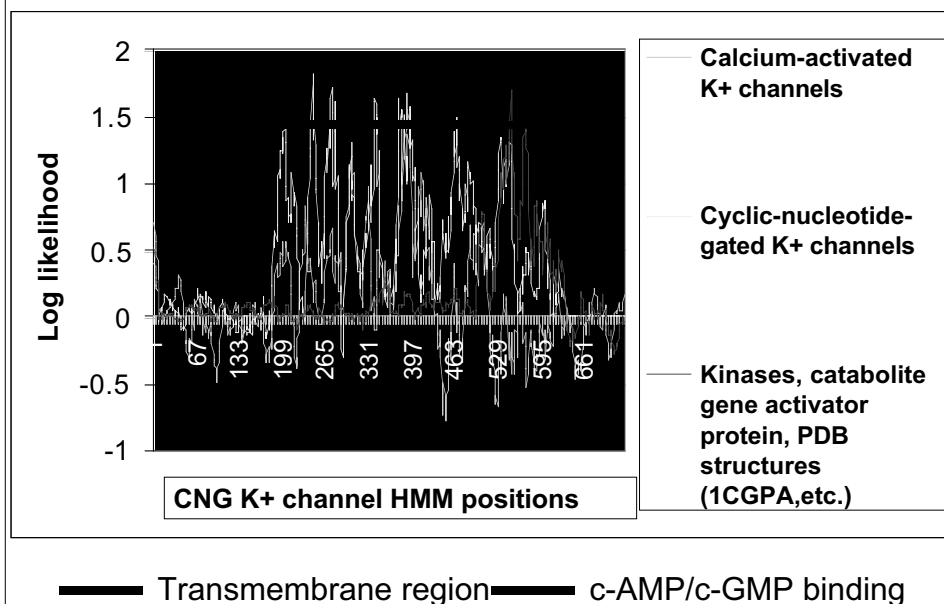


Nck_human
(adapter protein)
(PFAM domain identification shown)

Knowledge of domain structure is critical for correct function prediction



Analysis of clusters plots



Given a sequence...

NCBI Sequence Viewer - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=15236051&opt=GenPept

NCBI Protein

Search: Protein for [Go] [Clear]

Display: default [Save] [Text] [Add to Clipboard]

1: NP_195692: hypothetical prot. [gi.15236051] BLink, Nucleotide, OMIM, Relat

LOCUS NP_195692 427 aa linear PLN 10-JAN-2002

DEFINITION hypothetical protein [Arabidopsis thaliana].

ACCESSION NP_195692

VERSION NP_195692.1 GI:15236051

FEATURES

ORIGIN

1: NP_195692: hypothetical prot. [gi.15236051] BLink, Nucleotide, OMIM, Relat

LOCUS NP_195692 427 aa linear PLN 10-JAN-2002

DEFINITION hypothetical protein [Arabidopsis thaliana].

ACCESSION NP_195692

VERSION NP_195692.1 GI:15236051

FEATURES

ORIGIN

Try PFAM

Pfam - HMM Search (Saint Louis) - Microsoft Internet Explorer

Address: <http://pfam.sussex.ac.uk/hmmsearch.shtml>

Washington University in St. Louis

Pfam - HMM Search

Analyze a query sequence using the Pfam HMM database

[Pfam (2. Local)] [Pfam (Cambridge)] [Pfam (Stockholm)] [Pfam (Trinity)] [HMMER] [WebHMM] [Oncology]

[Home] [Tutorial] [FAQ] [Data] [Browse Pfam] [Keyword search] [Taxonomy search] [Contact Us] [Help]

Analyze a query sequence by searching Pfam HMMs

Protein sequence query: Cut and paste your sequence here. FASTA format or raw sequence are acceptable.

On: Select the query sequence file you wish to use [Browse...]

[Submit Query] [Reset]

More advanced options

Search mode: [global and local alignments merged]

Cutoff strategy: [Pfam gathering threshold (GA)] [E-value 1.0]

Comments, questions, flames? Email: pfam@genetics.wustl.edu

Last modified: Thursday, 11-Apr-2002 18:06:35 CDT

Gather homologs with PSI-BLAST

NCBI **BLAST** - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/blast/blast.cgi

NCBI *formatting* **BLAST**
Nucleotide Protein Translations Retrieve results for an RCL

Your request has been successfully submitted and put into the Blast Queue.

Query = gl|523605.1|pe|NP_195692.1|hypothetical protein [Arabidopsis thaliana] (427 letters)

Hit the button to **See conserved domains from CDD**

The request ID is

Format! or **Reset**

The results are estimated to be ready in 1 minutes 20 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

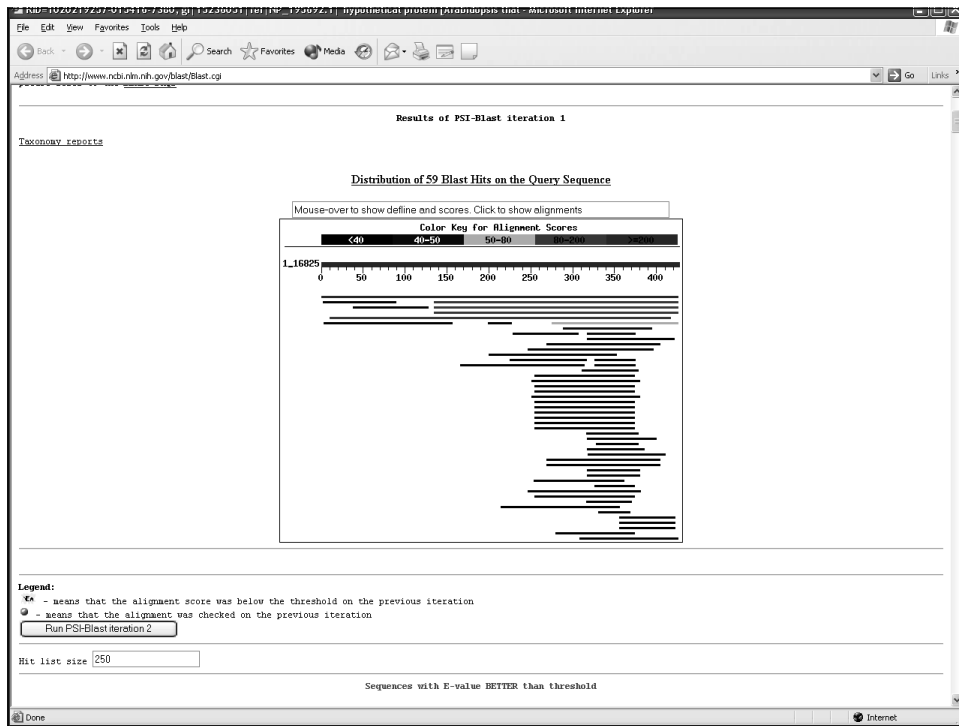
Show ☒ Graphical Overview ☒ Linkout ☒ to EBI ☐ Alignment in HTML

Number of Descriptions Alignments

Alignment view Pairwise

Format for PSI-BLAST ☒ with inclusion threshold

[illegible]



Sequences with E-value BETTER than threshold

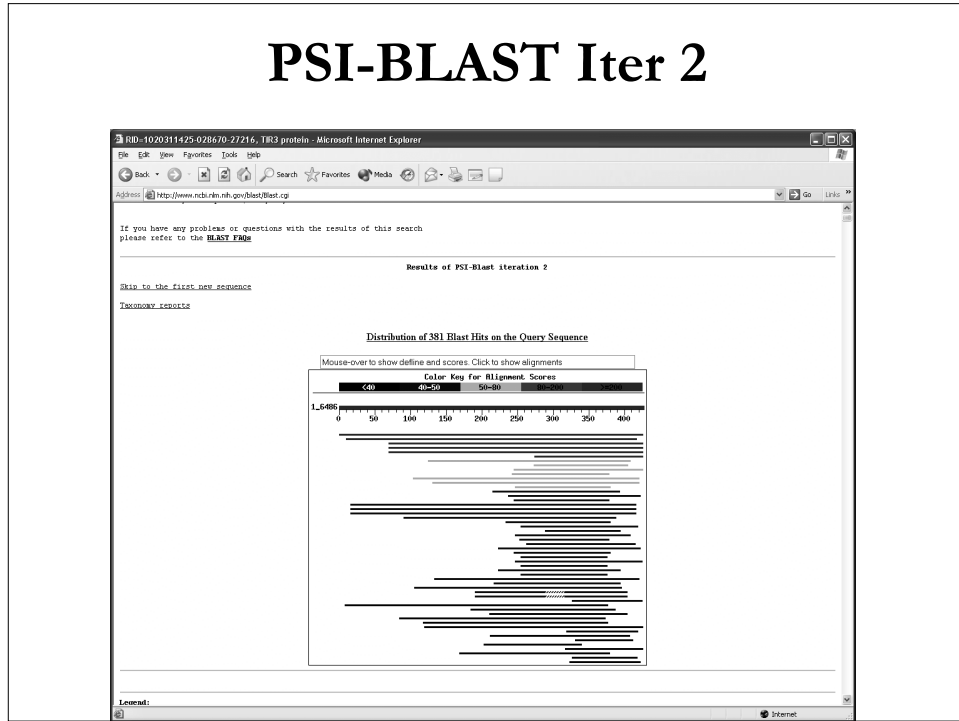
Sequences producing significant alignments:	Score (bits)	E Value
✖ <input checked="" type="checkbox"/> gi15236051 ref NP_195692.1 (NM_120145) hypothetical protein [A...	848	0.0
✖ <input checked="" type="checkbox"/> gi117315171 gb AAH14164.1 AAH14164 (BC014164) Unknown (protein f...	134	3e-30
✖ <input checked="" type="checkbox"/> gi120127545 ref NP_057114.2 (NM_016030) CGI-87 protein [Homo sa...	132	1e-29
✖ <input checked="" type="checkbox"/> gi14929643 gb AAD34082.1 AF151845.1 (AF151845) CGI-87 protein [H...	131	1e-29
✖ <input checked="" type="checkbox"/> gi110726916 gb AAF51639.2 (AE003592) CGI1396 gene product [Dros...	115	6e-25
✖ <input checked="" type="checkbox"/> gi112805517 gb AAH02235.1 AAH02235 (BC002235) Similar to CGI-87 ...	67	3e-10

Run PSI-Blast iteration 2

Sequences with E-value WORSE than threshold

<input type="checkbox"/> gi120091147 ref NP_617222.1 (NC_003552) hypothetical protein (m...	41	0.028
<input type="checkbox"/> gi115895668 ref NP_349017.1 (NC_003030) TPR repeats containing ...	40	0.053
<input type="checkbox"/> gi120090471 ref NP_616546.1 (NC_003552) TPR-domain containing p...	40	0.065
<input type="checkbox"/> gi119527172 ref NP_598713.1 (NM_133952) expressed sequence AW53...	39	0.15
<input type="checkbox"/> gi13599376 gb AAC62682.1 (AF083071) hypothetical protein 02 [Ce...	38	0.17
<input type="checkbox"/> gi115895716 ref NP_349065.1 (NC_003030) SoxR family transcripti...	37	0.27
<input type="checkbox"/> gi1872116 emb CAA56165.1 (X79770) sti (stress inducible protein...	37	0.41
<input type="checkbox"/> gi12129844 pir 1356658 stress-induced protein stil - soybean	37	0.42
<input type="checkbox"/> gi117563052 ref NP_503322.1 (NM_070921) R0912.3.p [Caenorhabdi...	37	0.51
<input type="checkbox"/> gi113324592 gb AAK18799.1 AF305607.1 (AF305607) LMP1 [Borrelia b...	36	0.74
<input type="checkbox"/> gi120093095 ref NP_619170.1 (NC_003552) TPR-domain containing p...	35	1.1

PSI-BLAST Iter 2



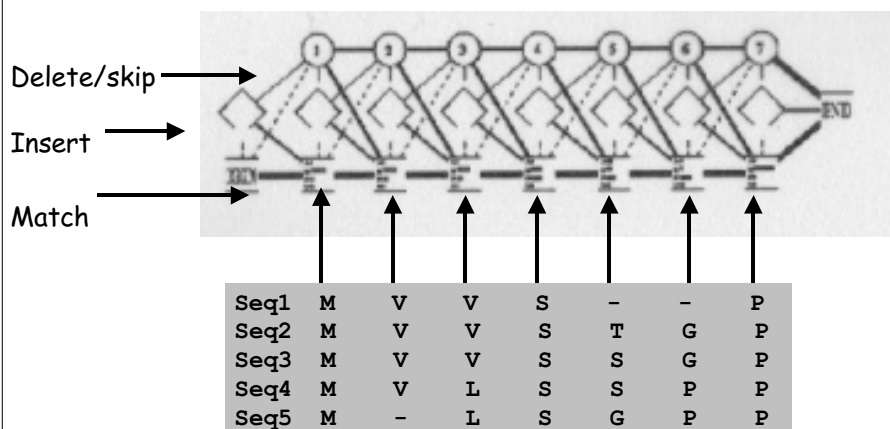
Alignment of close homologs

TIR3 = 15236051

(Using the Belvu alignment viewer)

iter1.fa			
gi 10726916 gb AAAF51639.2	1	mlrflgkaaaakqvlnadsveqsfvqlkqlisornraavdlgrlltahegagugksiltshttdslqlwfulal..	1
gi 17028344 gb AAH17475.1 AAH17475	1	mlrflgkaaaakqvlnadsveqsfvqlkqlisornraavdlgrlltahegagugksiltshttdslqlwfulal..	1
gi 17315171 gb AAH14164.1 AAH14164	1	mlrflgkaaaakqvlnadsveqsfvqlkqlisornraavdlgrlltahegagugksiltshttdslqlwfulal..	1
gi 17705797 ref INP_057114.1	1	mlrflgkaaaakqvlnadsveqsfvqlkqlisornraavdlgrlltahegagugksiltshttdslqlwfulal..	1
gi 10726916 gb AAAF51639.2	81	lwelprenepgklimpigldtldadpisvavtqhlgeaelsyrkilrvddvtqdengrltlihgacrtavntlgrilt	81
gi 10726916 gb AAAF51639.2	161	iygagugrsgqpkahspshlqlwfulal..-----	161
gi 15236051 ref INP_195692.1	1MVSIGKTTQIQRPNOVTSTMTTESSPANDPDPSETRPEFNPSSTDSTAM	1
gi 15236051 ref INP_195692.1	52	AESTDAGEPTAFELASSQVTSVADLPPERFNSLDELTHDLGSLHELSTRGSWQAILEKISQARALFLTKPHEHLYTLY	52
gi 10726916 gb AAAF51639.2	190	----LAKLGEFELLNRAEPEGOITSPDVEYDEYPERVYNGKSGSTACFSFRLLRLAEPTLYLGGKPHVALDRLSLHVTTRRE	190
gi 15236051 ref INP_195692.1	132	QVMALAKLRSDERSHELNSLHDFDGHYRVECFEYVPHRGSMWPFSLRLVYALPTKLGRNDEGLDRLYVLDFVRD	132
gi 17028344 gb AAH17475.1 AAH17475	79	----LVKLGLFQNAEMEFEPFGNLDQPDLYEYYPHYVPGRRGSMWPFSLRLVYALPTKLGRNDEGLDRLYVLDFVRD	79
gi 17315171 gb AAH14164.1 AAH14164	79	----LVKLGLFQNAEMEFEPFGNLDQPDLYEYYPHYVPGRRGSMWPFSLRLVYALPTKLGRNDEGLDRLYVLDFVRD	79
gi 17705797 ref INP_057114.1	79	----LVKLGLFQNAEMEFEPFGNLDQPDLYEYYPHYVPGRRGSMWPFSLRLVYALPTKLGRNDEGLDRLYVLDFVRD	79
gi 10726916 gb AAAF51639.2	266	IKKHYLSLHNAK.....EEFWDRRCERVLHSTINGCLHKKKFSMIDIMEGLLNRSN.LekedRSELYS	266
gi 15236051 ref INP_195692.1	212	PIREKESQSLK.....SVEIMKKRETFVWNCLLGFHLGHKEGCVSLDLMK.ELINRDP.L....DPVLTISK	212
gi 17028344 gb AAH17475.1 AAH17475	155	TLANLEQGLAEDgmssvtqgrgaSIRLMRSRLGRVMYSHANGLLLMKDYVLAVDAYH.SVIRKYPaQ....EPQLLSG	155
gi 17315171 gb AAH14164.1 AAH14164	155	TLANLEQGLAEDgmssvtqgrgaSIRLMRSRLGRVMYSHANGLLLMKDYVLAVDAYH.SVIRKYPaQ....EPQLLSG	155
gi 17705797 ref INP_057114.1	155	TLANLEQGLAEDgmssvtqgrgaSIRLMRSRLGRVMYSHANGLLLMKDYVLAVDAYH.SVIRKYPaQ....EPQLLSG	155
gi 10726916 gb AAAF51639.2	332	WGRTYLDGTGTFGAEQKI-AVSRRRLREINSAPDRLD-----VDKGLIAVAKNDPFAWVIFOKAHLDTGHTILNN	332
gi 15236051 ref INP_195692.1	273	LGSVMQFGDVEGAKTTDFRVEKMLNEKSNGLLNEIQFNILVGRNKALVYVAKDYVSARVEYDKCERDINSITAVNN	273
gi 17028344 gb AAH17475.1 AAH17475	230	IGRISLOIGDKITAEKYFDVEKVTQK-----LDGLGKINVLNNSAFHLGQNNFAEARRFFTEILRDPNRAVANN	230
gi 17315171 gb AAH14164.1 AAH14164	230	IGRISLOIGDKITAEKYFDVEKVTQK-----LDGLGKINVLNNSAFHLGQNNFAEARRFFTEILRDPNRAVANN	230
gi 17705797 ref INP_057114.1	230	IGRISLOIGDKITAEKYFDVEKVTQK-----LDGLGKINVLNNSAFHLGQNNFAEARRFFTEILRDPNRAVANN	230
gi 10726916 gb AAAF51639.2	404	HCCLLYLQGLKDSLRQLEAMVQDDPRYLHESVLFNLTTMYELESSRSMOKKQALLEAVAGKEGDSFNTQC-----IKLa	404
gi 15236051 ref INP_195692.1	363	KALCLLYLQGLKDSLRQLEAMVQDDPRYLHESVLFNLTTMYELESSRSMOKKQALLEAVAGKEGDSFNTQC-----IKLa	363
gi 17028344 gb AAH17475.1 AAH17475	304	ARVCLLYLQGLKDSLRQLEAMVQDDPRYLHESVLFNLTTMYELESSRSMOKKQALLEAVAGKEGDSFNTQC-----IKLa	304
gi 17315171 gb AAH14164.1 AAH14164	304	ARVCLLYLQGLKDSLRQLEAMVQDDPRYLHESVLFNLTTMYELESSRSMOKKQALLEAVAGKEGDSFNTQC-----IKLa	304
gi 17705797 ref INP_057114.1	304	ARVCLLYLQGLKDSLRQLEAMVQDDPRYLHESVLFNLTTMYELESSRSMOKKQALLEAVAGKEGDSFNTQC-----IKLa	304
gi 10726916 gb AAAF51639.2	475	eiclkltatin. 484	475

Protein fold prediction using HMMs



Using UCSC SAM software: 'modelfromalign foo -alignfile <msa>' will create an HMM, 'foo.mod'. Failed to find homologous PDB structures with the HMM...

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:	Score (bits)	E Value
<input checked="" type="checkbox"/> gi15236051 ref NP_195692.1 (NM_120145) hypothetical protein [A...	848	0.0
<input checked="" type="checkbox"/> gi117315171 gb AAH14164.1 AAH14164 (BC014164) Unknown (protein f...	134	3e-30
<input checked="" type="checkbox"/> gi120127545 ref NP_057114.2 (NM_016030) CGI-87 protein [Homo sa...	132	1e-29
<input checked="" type="checkbox"/> gi14929643 gb AAD34082.1 AF151845.1 (AF151845) CGI-87 protein [H...	131	1e-29
<input checked="" type="checkbox"/> gi110726916 gb AAFS1639.2 (AF003592) CGI1396 gene product [Dros...	115	6e-25
<input checked="" type="checkbox"/> gi112805517 gb AAH02235.1 AAH02235 (BC002235) Similar to CGI-87 ...	67	3e-10

Run PSI-Blast iteration 2

Sequences with E-value WORSE than threshold

<input type="checkbox"/> gi120091147 ref NP_617222.1 (NC_003552) hypothetical protein (m...	41	0.028
<input type="checkbox"/> gi115895668 ref NP_349017.1 (NC_003030) TPR repeats containing ...	40	0.053
<input type="checkbox"/> gi120090471 ref NP_616546.1 (NC_003552) TPR-domain containing p...	40	0.065
<input type="checkbox"/> gi119527172 ref NP_598713.1 (NM_133952) expressed sequence AW53...	39	0.15
<input type="checkbox"/> gi13599376 gb AAC62682.1 (AF083071) hypothetical protein 02 [Ce...	38	0.17
<input type="checkbox"/> gi115895716 ref NP_349065.1 (NC_003030) SoxR family transcripti...	37	0.27
<input type="checkbox"/> gi1872116 emb CAA56165.1 (X79770) sti (stress inducible protein...	37	0.41
<input type="checkbox"/> gi12129844 pir I1856658 stress-induced protein stil - soybean	37	0.42
<input type="checkbox"/> gi117563052 ref NP_503322.1 (NM_070921) R09E12.3-p [Caenothabdi...	37	0.51
<input type="checkbox"/> gi113324592 gb AAK18799.1 AF305607.1 (AF305607) LMP1 [Borrelia b...	36	0.74
<input type="checkbox"/> gi120093095 ref NP_619170.1 (NC_003552) TPR-domain containing p...	35	1.1

Checking the top remote homolog

Do the two sequences agree at critical positions?
Is the alignment local, or global?

```
>gi|20091147|ref|NP_617222.1| (NC_003552) hypothetical protein (multi-domain) [Methanosarcina  
acetivorans str. C2A]  
gi|19916251|gb|AAM05702.1| (AE010918) hypothetical protein (multi-domain) [Methanosarcina  
acetivorans str. C2A]  
Length = 463  
  
Score = 40.8 bits (94), Expect = 0.028  
Identities = 34/112 (30%), Positives = 49/112 (43%), Gaps = 7/112 (6%)  
  
Query: 290 FDRVEKMLNEGKSNGLLNEIQFNNLVGR---NKALVYVVAQDYVSAVREYDKQIERDQSD 346  
          FD V K++N      G +  +F+N +      N + Y A Y A      YDK I D +  
Sbjct: 318 FDEVLKLVNAGLTG-MKLTEFSNSISDDWYMMGVYEQASRYDEAANCYDKAIRIDPLN 376  
  
Query: 347 IIAVNNRHALCLMYLRDLSDAIKVMESALERVPT---AALNESLVVNLCSMYE 395  
          A NN+ + L      D+IK E A E P+  A N+ L ++      YE  
Sbjct: 377 AKAVNNRQVILAIIEEKYEDSIKYFEVATELKPSMVDANFNKGLALSRIQKYE 428
```

Looking at 20091147

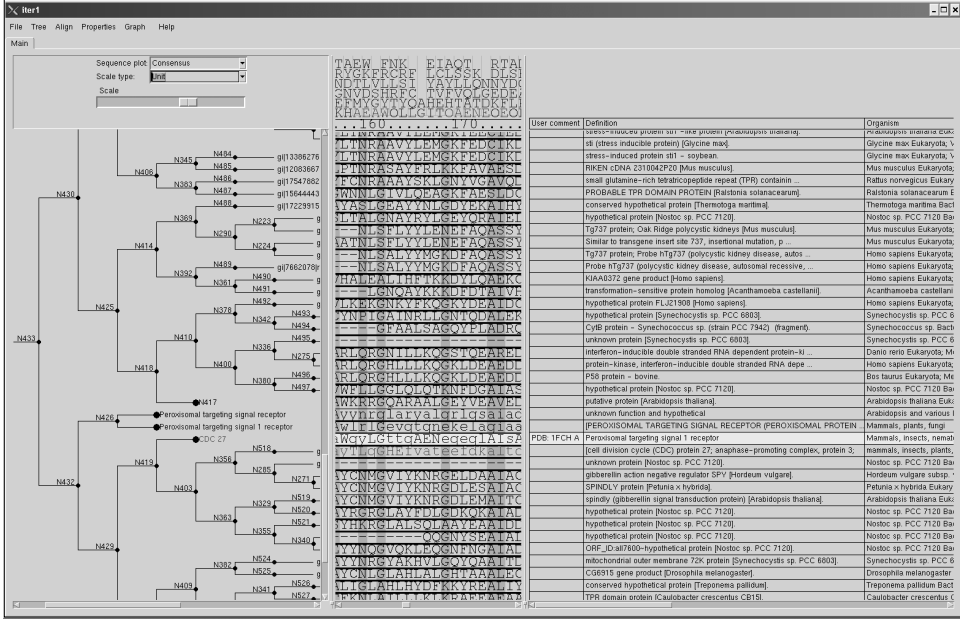
The screenshot shows the NCBI Sequence Viewer interface in a Microsoft Internet Explorer browser window. The address bar displays the URL: http://www.ncbi.nlm.nih.gov/jentrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=20091147&dopt=GenPept. The NCBI logo is visible at the top left, and the 'Protein' tab is selected. The search bar contains '20091147'. Below the search bar, there are buttons for 'Display', 'default', 'Save', 'Text', 'Add to Clipboard', 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The main content area displays the following information:

- LOCUS** NP_617222 463 aa linear BCT 09-APR-2002
- DEFINITION** hypothetical protein (multi-domain) [Methanosarcina acetivorans str. C2A].
- ACCESSION** NP_617222
- PID** g20091147
- VERSION** NP_617222.1 GI:20091147
- DBSOURCE** REFSEQ: accession NC_003552.1
- SOURCE** Methanosarcina acetivorans str. C2A.
- ORGANISM** Methanosarcina acetivorans str. C2A
- REFERENCE** 1 (residues 1 to 463)
- AUTHORS** Galagan, J.E., Huhsmann, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Sainov, S., Atmoor, D., Brown, A., Allen, H., Naylor, J., Stange-Thomann, M., DeArelano, R., Johnson, R., Linton, L., McDermott, P., McDermott, K., Talaas, J., Tittell, A., Ye, W., Zinner, A., Barber, R.D., Cann, I., Graham, D.E., Grubbs, D.A., Guss, A., Hedderich, R., Ingram-Smith, C., Kuettnier, C.H., Krzycki, J.A., Leigh, J.A., Li, W., Liu, J., Mukhopadhyay, D., Reeve, J.H., Smith, R., Springer, T.A., Umayak, L.A., White, G., White, R.H., de Macario, E.C., Ferry, J.G., Jurell, R.F., Ling, H., Macario, A.J.L., Paulsen, I., Pritchett, M., Sowers, K.R., Swanson, R.V., Zinder, S.H., Lander, E., Metcalf, W.W. and Birren, B.
- TITLE** The Genome of M. acetivorans Reveals Extensive Metabolic and Physiological Diversity
- JOURNAL** Genome Res. 12 (4), 532-542 (2002)
- MEDLINE** 21529760
- PMID** 11927798

Gather homologs, align...
try to identify critical positions

Extending the 20091147 alignment

Examining the tree: there's a structure!



Is the PDB structure 1FCH A homologous?

[illegible]

Close-up of joint alignment

1FCHA homologs above red line; TIR3 homologs below red line.
Sequences are all aligned to a subfamily HMM for 1FCH cluster.

1FCHA
TIR3

```

QACEILRDWLRYPFA.YahlvtpaeeGAGGAGLGPSKRILGslsdsLFLEVKEFLAAVRLDPT.....
EAIALFKKALVEHRN.P.....DT-----LAKLT.....ACEKEKEKFEIEAYLDPE.....
EAIKDLKKSISLDAS.Q.....PDYHNVLGSVYEDMG.....LVOKATQEFATAIKIEND.....
VSLDLMKELINRDPL.D.....PVLISKLGSVQMDFG.....DVEGAKITTFDRVEKMLNEgksngline
VSLDLMKELINRDPL.D.....PVLISKLGSVQMDFG.....DVEGAKITTFDRVEKMLNEgksngline
LAVEAYHSHVTKYYPEqE.....PQLLSGIGRISLQIG.....DIKTAEKYFQDVEKVTQK.....
LAVDAYHSHVTKYYPEqE.....PQLLSGIGRISLQIG.....DIKTAEKYFQDVEKVTQK.....
LAVEAYHSHVTKYYPEqE.....PQLLSGIGRISLQIG.....DIKTAEKYFQDVEKVTQK.....
.....IDPDVQcGLGVLFNLSGEYDKAVDCFTAALSVRPNQYLLWNKLGATLANGNQSEEAAYRRALQLPGY
.....ALQKKE..EGNTFFKSDKEPEAVEAYTEAIKRNPDHTTYSNRAAYLKLGAYSQALADAEKICSLKPEF
.....PDYYYN..RGNAYWKKGDVDKALEDYSKAADLDSTQIFVYKKYEALMNLGRNLNEALATIEKAIVPAN
.....KCNFEESQCPVMEQVNYESKATQIDFANSVLYSNPSQCFQOMKYYKDALDDQKCTSTKPNL
.....NLVGRN..KALVYVVAKDYSAVREYDKCIERDNSIIAVNNKALCLMYLRDLSDAIKVMESALERV--
.....--LVD..KGLIAVAKNDFPEAYVIAQKALHLDGTNTMILNNMGVCLLYAGKLDAINLYERAINLNQ-
.....NLVGRN..KALVYVVAKDYSAVREYDKCIERDNSIIAVNNKALCLMYLRDLSDAIKVMESALERV--
1qgkIMVLMN..SAFLHLGQNNFAEAHRFFETILRMDPRNAVANNNAAVCLLYLGKLDKSLRQLEAMVQQDPRH
1qgkIMVLMN..SAFLHLGQNNFAEAHRFFETILRMDPRNAVANNNAAVCLLYLGKLDKSLRQLEAMVQQDPRH
1qgkIMVLMN..SAFLHLGQNNFAEAHRFFETILRMDPRNAVANNNAAVCLLYLGKLDKSLRQLEAMVQQDPRH

```

SCOP: Superfamily: Tetratricopeptide repeat (TPR) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print Mail News RSS

Address <http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.b.bce.i.a.html> Go Links

Structural Classification of Proteins

?

Superfamily: Tetratricopeptide repeat (TPR)

Lineage:

1. Root: scop
2. Class: All alpha proteins
3. Fold: alpha-alpha superhelix
multihelical, 2 (curved) layers: alpha/alpha, right-handed superhelix
4. Superfamily: Tetratricopeptide repeat (TPR)

Families:

1. Tetratricopeptide repeat (TPR) (7)
 1. Protein phosphatase 5
 1. Human (*Homo sapiens*) (1) ▾
 2. Hop
 1. Human (*Homo sapiens*) (2) ▾
 3. Vesicular transport protein sec17
 1. Baker's yeast (*Saccharomyces cerevisiae*) (1) ▾
 4. Neutrophil cytosolic factor 2 (NCF-2, p67-phox)
 1. Human (*Homo sapiens*) (2) ▾
 5. Peroxin per2 (peroxisomal targeting signal 1 (PTS1) receptor)
 1. Human (*Homo sapiens*) (1) ▾
 2. *Trypanosoma brucei* (1) ▾
 6. Cyclophilin 40
 1. Cow (*Bos taurus*) (2) ▾

Enter search key: Search

df=49452 [a.118.8] Internet

The
TPR
SCOP
Super
family

Closing notes

- Homolog detection is just the first step
 - Attention to domain structure is critical
 - Phylogenetic inference is key to functional analysis
 - Thanks to:
 - David Haussler, Kevin Karplus, Chris Sander
 - Barbara Baker, Brian Staskawicz, Richard Michelson (Plant Biologist collaborators)
- kimmen@uclink.berkeley.edu